HUMAN DNA MISMATCH REPAIR PROTEINS

[0001] This application is a divisional of, and claims the benefit of priority under 35 U.S.C. § 120 of, U.S. Patent Application Serial No. 08/468,024 filed June 6, 1995, which is a continuation-in-part of, and claims the benefit of priority under 35 U.S.C. § 120 of, International Application No. PCT/US95/01035 filed January 25, 1995; U.S. Patent Application Serial No. 08/468,024 filed June 6, 1995 is also a continuation-in-part of, and claims the benefit of priority under 35 U.S.C. § 120 of, U.S. Patent Application Serial No. 08/294,312 filed August 23, 1994, which is a continuation-in-part of, and claims the benefit of priority under 35 U.S.C. § 120 of, U.S. Patent Application Serial No. 08/210,143 filed March 16, 1994, which is a continuation-in-part of, and claims the benefit of priority under 35 U.S.C. § 120 of, U.S. Patent Application Serial No. 08/187,757 filed January 27, 1994; and this application is also a divisional of, and claims the benefit of priority under 35 U.S.C. § 120 of, U.S. Patent Application Serial No. 08/465,769 filed June 6, 1995, which is a continuation-in-part of, and claims the benefit of priority under 35 U.S.C. § 120 of, U.S. Patent Application Serial No. 08/294,312 filed August 23, 1994, which is a continuation-in-part of, and claims the benefit of priority under 35 U.S.C. § 120 of, U.S. Patent Application Serial No. 08/210,143 filed March 16, 1994, which is a continuation-in-part of, and claims the benefit of priority under 35 U.S.C. § 120 of, U.S. Patent Application Serial No. 08/187,757 filed January 27, 1994; and this application is also a continuation-in-part of, and claims the benefit of priority under 35 U.S.C. § 120 of, U.S. Patent Application Serial No. 08/294,312 filed August 23, 1994, which is a continuation-in-part of, and claims the benefit of priority under 35 U.S.C. § 120 of, U.S. Patent Application Serial No. 08/210,143 filed March 16, 1994, which is a continuation-in-part of, and claims the benefit of priority under 35 U.S.C. § 120 of, U.S. Patent Application Serial No. 08/187,757 filed January 27, 1994; and this application is also a continuation-in-part of, and claims the benefit of priority under 35 U.S.C. § 120 of, U.S. Patent Application Serial No. 08/210,143 filed March 16,

1994, which is a continuation-in-part of, and claims the benefit of priority under 35 U.S.C. § 120 of, U.S. Patent Application Serial No. 08/187,757 filed January 27, 1994; and this application is also a continuation-in-part of, and claims the benefit of priority under 35 U.S.C. § 120 of, U.S. Patent Application Serial No. 08/187,757 filed January 27, 1994. Each of the aforementioned U.S. and International patent applications are hereby incorporated by reference in their entireties.

Field of the Invention

[0002] This invention relates to newly identified polynucleotides, polypeptides encoded by such polynucleotides, the use of such polynucleotides and polypeptides, as well as the production of such polynucleotides and polypeptides. More particularly, the polypeptides of the present invention are human homologs of the prokaryotic mutL*A* gene and are hereinafter referred to as hMLH1, hMLH2 and hMLH3.

[0003] In both ~~prolaryotes~~ **prokaryotes** and eukaryotes, the DNA mismatch repair gene plays a prominent role in the correction of errors made during DNA replication and genetic recombination. The E.coli methyl-directed DNA mismatch repair system is the best understood DNA mismatch repair system to date. In E.coli, this repair pathway involves the products of the mutator genes mutS, mutL, mutH, and uvrD. Mutants of any one of these genes will reveal a mutator phenotype. MutS is a DNA mismatch-binding protein which initiates this repair process, uvrD is a DNA helicase and MutH is a latent endonuclease that incises at the unmethylated strands of a hemi-methylated GATC sequence. MutL protein is believed to recognize and bind to the mismatch-DNA-MutS-MutH complex to enhance the endonuclease activity of MutH protein. After the unmethylated DNA strand is cut by the MutH, single-stranded DNA-binding protein, DNA polymerase III, exonuclease I and DNA ligase are required to complete this repair process (Modrich P., Annu. Rev. Genetics, 25:229-53 (1991)).

[0004] Elements of the E.coli MutLHS system appears to be conserved during evolution in prokaryotes and eukaryotes. Genetic study analysis suggests that Saccharomyces cerevisiae has a mismatch repair system similar to the bacterial MutLHS system. In S. cerevisiae, at least two MutL homologs, PMS1 and MLH1, have been

reported. Mutation of either one of them leads to a mitotic mutator phenotype (Prolla et al, Mol. Cell. Biol. 14:407-415 (1994)). At least three MutS homologs have been found in S.cerevisiae, namely MSH1, MSH2, and MSH3. Disruption of the MSH2 gene affects nuclear mutation rates. Mutants in S. cerevisae, MSH2, PMS1, and MLH1 have been found to exhibit increased rates of expansion and contraction of dinucleotide repeat sequences (Strand et al., Nature, 365:274-276 (1993)).

[0005] It has been reported that a number of human tumors such as lung cancer, prostate cancer, ovarian cancer, breast cancer, colon cancer and stomach cancer show instability of repeated DNA sequences (Han et al., Cancer, 53:5087-5089 (1993); Thibodeau et al., Science 260:816-819 (1993); Risinger et al., Cancer 53:5100-5103 (1993)). This phenomenon suggests that lack of the DNA mismatch repair is probably the cause of these tumors.

[0006] Little was known about the DNA mismatch repair system in humans until recently, the human homolog of the MutS gene was cloned and found to be responsible for hereditary nonpolyposis colon cancer (HNPCC), (Fishel et al., Cell, 75:1027-1038 (1993) and Leach et al., Cell, 75:1215-1225 (1993)). HNPCC was first linked to a locus at chromosome 2p16 which causes dinucleotide instability. It was then demonstrated that a DNA mismatch repair protein (MutS) homolog was located at this locus, and that C-->T transitional mutations at several conserved regions were specifically observed in HNPCC patients. Hereditary nonpolyposis colorectal cancer is one of the most common hereditable diseases of man, affecting as many as one in two hundred individuals in the western world.

[0007] It has been demonstrated that hereditary colon cancer can result from mutations in several loci. Familial adenomatosis polyposis coli (APC), linked to a gene on chromosome 5, is responsible for a small minority of hereditary colon cancer. Hereditary colon cancer is also associated with Gardner's syndrome, Turcot's syndrome, Peutz-Jaeghers syndrome and juvenile polyposis coli. In addition, hereditary nonpolyposis colon cancer may be involved in 5% of all human colon cancer. All of the different types of familial colon cancer have been shown to be transmitted by a dominant autosomal mode of inheritance.

[0008] In addition to localization of HNPCC, to the short arm of chromosome 2, a second locus has been linked to a pre-disposition to HNPCC (Lindholm, et al., Nature Genetics, 5:279-282 (1993)). A strong linkage was demonstrated between a polymorphic marker on the short arm of chromosome 3 and the disease locus.

[0009] This finding suggests that mutations on various DNA mismatch repair proteins probably play crucial roles in the development of human hereditary diseases and cancers.

[0010] HNPCC is characterized clinically by an apparent autosomal dominantly inherited predisposition to cancer of the colon, endometrium and other organs. (Lynch, H.T. et al., Gastroenterology, 104:1535-1549 (1993)). The identification of markers at 2p16 and 3p21-22 which were linked to disease in selected HNPCC kindred unequivocally established its mendelian nature (Peltomaki, P. et al., Science, 260:810-812 (1993)). Tumors from HNPCC patients are characterized by widespread alterations of simple repeated sequences (microsatellites) (Aaltonen, L.A., et al., Science, 260:812-816 (1993)). This type of genetic instability was originally observed in a subset (12 to 18% of sporadic colorectal cancers (Id.). Studies in bacteria and yeast indicated that a defect in DNA mismatch repair genes can result in a similar instability of microsatellites (Levinson, G. and Gutman, G.A., Nuc. Acids Res., 15:5325-5338 (1987)), and it was hypothesized that deficiency in mismatched repair was responsible for HNPCC (Strand, M. et al., Nature, 365:274-276 (1993)). Analysis of extracts from HNPCC tumor cell lines showed mismatch repair was indeed deficient, adding definitive support to this conjecture (Parsons, R.P., et al., Cell, 75:1227-1236 (1993)). As not all HNPCC kindred can be linked to the same loci, and as at least three genes can produce a similar phenotype in yeast, it seems likely that other mismatch repair genes could play a role in some cases of HNPCC.

Summary of the Invention

[0011] hMLH1 is most homologous to the yeast mutL-homolog yMLH1 while hMLH2 and hMLH3 have greater homology to the yeast mutL-homolog yPMS1 (hMLH2 and hMLH3 due to their homology to yeast PMS1 gene are sometimes referred to in the literature as hPMS1 and hPMS2). In addition to hMLH1, both the hMLH2 gene on

chromosome 2q32 and the hMLH3 gene, on chromosome 7p22, were found to be mutated in the germ line of HNPCC patients. This doubles the number of genes implicated in HNPCC and may help explain the relatively high incidence of this disease.

[0012] In accordance with one aspect of the present invention, there are provided novel putative mature polypeptides which are hMLH1, hMLH2 and hMLH3, as well as biologically active and diagnostically or therapeutically useful fragments, analogs and derivatives thereof. The polypeptides of the present invention are of human origin.

[0013] In accordance with another aspect of the present invention, there are provided isolated nucleic acid molecules encoding such polypeptides, including mRNAs, DNAs, cDNAs, genomic DNA as well as biologically active and diagnostically or therapeutically useful fragments, analogs and derivatives thereof.

[0014] In accordance with still another aspect of the present invention there are provided nucleic acid probes comprising nucleic acid molecules of sufficient length to specifically hybridize to hMLH1, hMLH2 and hMLH3 sequences.

[0015] In accordance with yet a further aspect of the present invention, there is provided a process for producing such polypeptides by recombinant techniques which comprises culturing recombinant prokaryotic and/or eukaryotic host cells, containing an hMLH1, hMLH2 or hMLH3 nucleic acid sequence, under conditions promoting expression of said protein and subsequent recovery of said proteins.

[0016] In accordance with yet a further aspect of the present invention, there is provided a process for utilizing such polypeptide, or polynucleotide encoding such polypeptide, for therapeutic purposes, for example, for the treatment of cancers.

[0017] In accordance with another aspect of the present invention there is provided a method of diagnosing a disease or a susceptibility to a disease related to a mutation in the hMLH1, hMLH2 or hMLH3 nucleic acid sequences and the proteins encoded by such nucleic acid sequences.

[0018] In accordance with yet a further aspect of the present invention, there is provided a process for utilizing such polypeptides, or polynucleotides encoding such polypeptides, for in vitro purposes related to scientific research, synthesis of DNA and manufacture of DNA vectors.

[0019] These and other aspects of the present invention should be apparent to those skilled in the art from the teachings herein.

Brief Description of the Drawings

[0020] The following drawings are illustrative of embodiments of the invention and are not meant to limit the scope of the invention as encompassed by the claims.

[0021] Figure 1 illustrates the cDNA sequence and corresponding deduced amino acid sequence for the human DNA repair protein hMLH1. The amino acids are represented by their standard one-letter abbreviations. Sequencing was performed using a 373 Automated DNA sequencer (Applied Biosystems, Inc.). Sequencing accuracy is predicted to be greater than 97% accurate.

[0022] Figure 2 illustrates the cDNA sequence and corresponding deduced amino acid sequence of hMLH2. The amino acids are represented by their standard one-letter abbreviations.

[0023] Figure 3 illustrates the cDNA sequence and corresponding deduced amino acid sequence of hMLH3. The amino acids are represented by their standard one-letter abbreviations.

[0024] Figure 4. Alignment of the predicted amino acid sequences of S. cerevisiae PMS1 (yPMS1), with the hMLH2 and hMLH3 amino acid sequences using MACAW (version 1.0) program. Amino acid in conserved blocks are capitalized and shaded on the mean of their pair-wise scores.

[0025] Figure 5. Mutational analysis of hMLH2. (A) IVSP analysis and mapping of the transcriptional stop mutation in HNPCC patient CW. Translation of codons 1 to 369 (lane 1), codons 1 to 290 (lane 2), and codons 1 to 214 (lane 3). CW is translated from the cDNA of patient CW, while NOR was translated from the cDNA of a normal individual. The arrowheads indicate the truncated polypeptide due to the potential stop mutation. The arrows indicate molecular weight markers in kilodaltons. (B) Sequence analysis of CW indicates a C to T transition at codon 233 (indicated by the arrow). Lanes 1 and 3 are sequence derived from control patients; lane 2 is sequence derived from genomic DNA of

CW. The ddA mixes from each sequencing mix were loaded in adjacent lanes to facilitate comparison as were those for ddC, ddD, and ddT mixes.

[0026] Figure 6. Mutational analysis of hMLH3. (A) IVSP analysis of hMLH3 from patient GC. Lane GC is from fibroblasts of individual GC; lane GCx is from the tumor of patient GC; lanes NOR1 and 2 are from normal control individuals. FL indicates full-length protein, and the arrowheads indicate the germ line truncated polypeptide. The arrows indicate molecular weight markers in kilodaltons (B) PCR analysis of DNA from a patient GC shows that the lesion in present in both hMLH3 alleles in tumor cells. Amplification was done using primers that amplify 5', 3', or within (MID) the region deleted in the cDNA. Lane 1, DNA derived from fibroblasts of patient GC; lane 2, DNA derived from tumor of patient GC; lane 3, DNA derived from a normal control patient; lane 4, reactions without DNA template. Arrows indicate molecular weight in base pairs.

Detailed Description of the Invention

[0027] In accordance with an aspect of the present invention, there are provided isolated nucleic acids (polynucleotides) which encode for the mature polypeptides having the deduced amino acid sequence of Figures 1, 2 and 3 (SEQ ID NOS:2, 4 and 6) or for the mature polypeptides encoded by the cDNA of the clone deposited as ATCC Deposit No. 75649, 75651, 75650, deposited on January 25, 1994. **The address of the American Type Culture Collection (ATCC) Depository referred to herein is: American Type Culture Collection, 10801 University Boulevard, Manassas, Virginia 20110-2209.**

[0028] ATCC Deposit No. 75649 is a cDNA clone which contains the full length sequence encoding the human DNA repair protein referred to herein as hMLH1; ATCC Deposit No. 75651 is a cDNA clone containing the full length cDNA sequence encoding the human DNA repair protein referred to herein as hMLH2; ATCC Deposit No. 75650 is a cDNA clone containing the full length DNA sequence referred to herein as hMLH3.

[0029] Polynucleotides encoding the polypeptides of the present invention may be obtained from one or more libraries prepared from heart, lung, prostate, spleen, liver, gallbladder, fetal brain and testes tissues. The polynucleotides of hMLH1 were discovered from a human gallbladder cDNA library. In addition, six cDNA clones which are identical

to the hMLH1 at the N-terminal ends were obtained from human cerebellum, eight-week embryo, fetal heart, HSC172 cells and Jurket cell cDNA libraries. The hMLH1 gene contains an open reading frame of 756 amino acids encoding for an 85kD protein which exhibits homology to the bacterial and yeast *mutL* proteins. However, the 5' non-translated region was obtained from the cDNA clone obtained from the fetal heart for the purpose of extending the non-translated region to design the oligonucleotides.

[0030] The hMLH2 gene was derived from a human T-cell lymphoma cDNA library. The hMLH2 cDNA clone identified an open reading frame of 2,796 base pairs flanked on both sides by in-frame termination codons. It is structurally related to the yeast PMS1 family. It contains an open reading frame encoding a protein of ~~934~~ **932** amino acid residues. The protein exhibits the highest degree of homology to yeast PMS1 with 27% identity and 82 % similarity over the entire protein.

[0031] A second region of significant homology among the three PMS related proteins is in the carboxyl terminus, between codons 800 to 900. This region shares a 22% and 47% homology between yeast PMS1 protein and hMLH2 and hMLH3 proteins, respectively, while very little homology of this region was observed between these proteins, and the other yeast *mutL* homolog, yMLH1.

[0032] The hMLH3 gene was derived from a human endometrial tumor cDNA library. The hMLH3 clone identified a 2,586 base pair open reading frame. It is structurally related to the yPMS2 protein family. It contains an open reading frame encoding a protein of 862 amino acid residues. The protein exhibits the highest degree of homology to yPMS2 with 32% identity and 66% similarity over the entire amino acid sequence.

[0033] It is significant with respect to a putative identification of hMLH1, hMLH2 and hMLH3 that the GFRGEAL domain which is conserved in *mutL* homologs derived from *E. coli* is conserved in the amino acid sequences of , hMLH1, hMLH2 and hMLH3.

[0034] The polynucleotides of the present invention may be in the form of RNA or in the form of DNA, which DNA includes cDNA, genomic DNA, and synthetic DNA. The DNA may be double-stranded or single-stranded, and if single stranded may be the coding strand or non-coding (anti-sense) strand. The coding sequence which encodes the mature polypeptide may be identical to the coding sequence shown in Figures 1, 2 and 3 (SEQ ID

NO:1) or that of the deposited clone or may be a different coding sequence which coding sequence, as a result of the redundancy or degeneracy of the genetic code, encodes the same mature polypeptides as the DNA of Figures 1, 2 and 3 (SEQ ID NOS:2, 4 and 6) or the deposited cDNA(s).

[0035] The polynucleotides which encode for the mature polypeptides of Figures 1, 2 and 3 (SEQ ID NOS:2, 4 and 6) or for the mature polypeptides encoded by the deposited cDNAs may include: only the coding sequence for the mature polypeptide; the coding sequence for the mature polypeptide (and optionally additional coding sequence) and non-coding sequence, such as introns or non-coding sequence 5' and/or 3' of the coding sequence for the mature polypeptide.

[0036] Thus, the term "polynucleotide encoding a polypeptide" encompasses a polynucleotide which includes only coding sequence for the polypeptide as well as a polynucleotide which includes additional coding and/or non-coding sequence.

[0037] The present invention further relates to variants of the hereinabove described polynucleotides which encode for fragments, analogs and derivatives of the polypeptides having the deduced amino acid sequences of Figures 1, 2 and 3 (SEQ ID NOS:2, 4 and 6) or the polypeptides encoded by the cDNA of the deposited clones. The variants of the polynucleotides may be a naturally occurring allelic variant of the polynucleotides or a non-naturally occurring variant of the polynucleotides.

[0038] Thus, the present invention includes polynucleotides encoding the same mature polypeptides as shown in Figures 1, 2 and 3 (SEQ ID NOS:2, 4 and 6) or the same mature polypeptides encoded by the cDNA of the deposited clones as well as variants of such polynucleotides which variants encode for a fragment, derivative or analog of the polypeptides of Figures 1, 2 and 3 (SEQ ID NOS:2, 4 and 6) or the polypeptides encoded by the cDNA of the deposited clones. Such nucleotide variants include deletion variants, substitution variants and addition or insertion variants.

[0039] As hereinabove indicated, the polynucleotides may have a coding sequence which is a naturally occurring allelic variant of the coding sequence shown in Figures 1, 2 and 3 (SEQ ID NO:1, 3 and 5) or of the coding sequence of the deposited clones. As known in the art, an allelic variant is an alternate form of a polynucleotide sequence which

may have a substitution, deletion or addition of one or more nucleotides, which does not substantially alter the function of the encoded polypeptide.

[0040] The polynucleotides of the present invention may also have the coding sequence fused in frame to a marker sequence which allows for purification of the polypeptides of the present invention. The marker sequence may be, for example, a hexa-histidine tag supplied by a pQE-9 vector to provide for purification of the mature polypeptides fused to the marker in the case of a bacterial host, or, for example, the marker sequence may be a hemagglutinin (HA) tag when a mammalian host, e.g. COS-7 cells, is used. The HA tag corresponds to an epitope derived from the influenza hemagglutinin protein (Wilson, I., et al., Cell, 37:767 (1984)).

[0041] The term "gene" means the segment of DNA involved in producing a polypeptide chain; it includes regions preceding and following the coding region (leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons).

[0042] Fragments of the full length gene of the present invention may be used as a hybridization probe for a cDNA library to isolate the full length cDNA and to isolate other cDNAs which have a high sequence similarity to the gene or similar biological activity. Probes of this type preferably have at least 30 bases and may contain, for example, 50 or more bases. The probe may also be used to identify a cDNA clone corresponding to a full length transcript and a genomic clone or clones that contain the complete gene including regulatory and promotor regions, exons, and introns. An example of a screen comprises isolating the coding region of the gene by using the known DNA sequence to synthesize an oligonucleotide probe. Labeled oligonucleotides having a sequence complementary to that of the gene of the present invention are used to screen a library of human cDNA, genomic DNA or mRNA to determine which members of the library the probe hybridizes to.

[0043] The present invention further relates to polynucleotides which hybridize to the hereinabove-described sequences if there is at least 70%, preferably at least 90%, and more preferably at least 95% identity between the sequences. The present invention particularly relates to polynucleotides which hybridize under stringent conditions to the hereinabove-described polynucleotides. As herein used, the term "stringent conditions"

means hybridization will occur only if there is at least 95% and preferably at least 97% identity between the sequences. The polynucleotides which hybridize to the hereinabove described polynucleotides in a preferred embodiment encode polypeptides which either retain substantially the same biological function or activity as the mature polypeptide encoded by the cDNAs of Figure 1, 2 and 3 (SEQ ID NO:1, 3 and 5) or the deposited cDNA(s).

[0044] Alternatively, the polynucleotide may have at least 20 bases, preferably 30 bases, and more preferably at least 50 bases which hybridize to a polynucleotide of the present invention and which has an identity thereto, as hereinabove described, and which may or may not retain activity. For example, such polynucleotides may be employed as probes for the polynucleotide of SEQ ID NO:1, 3 and 5 for example, for recovery of the polynucleotide or as a diagnostic probe or as a PCR primer.

[0045] Thus, the present invention is directed to polynucleotides having at least a 70% identity, preferably at least 90% and more preferably at least a 95% identity to a polynucleotide which encodes the polypeptide of SEQ ID NOS:2, 4 and 6 as well as fragments thereof, which fragments have at least 30 bases and preferably at least 50 bases and to polypeptides encoded by such polynucleotides.

[0046] The deposit(s) referred to herein will be maintained under the terms of the Budapest Treaty on the International Recognition of the Deposit of Micro-organisms for purposes of Patent Procedure. These deposits are provided merely as convenience to those of skill in the art and are not an admission that a deposit is required under 35 U.S.C. §112. The sequence of the polynucleotides contained in the deposited materials, as well as the amino acid sequence of the polypeptides encoded thereby, are incorporated herein by reference and are controlling in the event of any conflict with any description of sequences herein. A license may be required to make, use or sell the deposited materials, and no such license is hereby granted.

[0047] The present invention further relates to polypeptides which have the deduced amino acid sequence of Figures 1, 2 and 3 (SEQ ID NOS:2, 4 and 6) or which have the amino acid sequence encoded by the deposited cDNA(s), as well as fragments, analogs and derivatives of such polypeptides.

[0048] The terms "fragment," "derivative" and "analog" when referring to the polypeptides of Figures 1, 2 and 3 (SEQ ID NOS:2, 4 and 6) or that encoded by the deposited cDNA(s), means polypeptides which retain essentially the same biological function or activity as such polypeptides. Thus, an analog includes a proprotein which can be activated by cleavage of the proprotein portion to produce an active mature polypeptide.

[0049] The polypeptides of the present invention may be a recombinant polypeptide, a natural polypeptide or a synthetic polypeptide, preferably a recombinant polypeptide.

[0050] The fragment, derivative or analog of the polypeptides of Figures 1, 2 and 3 (SEQ ID NOS:2, 4 and 6) or that encoded by the deposited cDNAs may be (i) one in which one or more of the amino acid residues are substituted with a conserved or non-conserved amino acid residue (preferably a conserved amino acid residue) and such substituted amino acid residue may or may not be one encoded by the genetic code, or (ii) one in which one or more of the amino acid residues includes a substituent group, or (iii) one in which the mature polypeptide is fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol). Such fragments, derivatives and analogs are deemed to be within the scope of those skilled in the art from the teachings herein.

[0051] The polypeptides and polynucleotides of the present invention are preferably provided in an isolated form, and preferably are purified to homogeneity.

[0052] The term "isolated" means that the material is removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or polypeptide, separated from some or all of the co-existing materials in the natural system, is isolated. Such polynucleotides could be part of a vector and/or such polynucleotides or polypeptides could be part of a composition, and still be isolated in that such vector or composition is not part of its natural environment.

[0053] The polypeptides of the present invention include the polypeptide of SEQ ID NOS:2, 4 and 6 (in particular the mature polypeptide) as well as polypeptides which have at least 70% similarity (preferably at least 70% identity) to the polypeptide of SEQ ID NOS:2, 4 and 6 and more preferably at least 90% similarity (more preferably at least 90%

identity) to the polypeptide of SEQ ID NOS:2, 4 and 6 and still more preferably at least 95% similarity (still more preferably at least 95% identity) to the polypeptide of SEQ ID NOS:2, 4 and 6 and also include portions of such polypeptides with such portion of the polypeptide generally containing at least 30 amino acids and more preferably at least 50 amino acids.

[0054] As known in the art "similarity" between two polypeptides is determined by comparing the amino acid sequence and its conserved amino acid substitutes of one polypeptide to the sequence of a second polypeptide.

[0055] Fragments or portions of the polypeptides of the present invention may be employed for producing the corresponding full-length polypeptide by peptide synthesis; therefore, the fragments may be employed as intermediates for producing the full-length polypeptides. Fragments or portions of the polynucleotides of the present invention may be used to synthesize full-length polynucleotides of the present invention.

[0056] The present invention also relates to vectors which include polynucleotides of the present invention, host cells which are genetically engineered with vectors of the invention and the production of polypeptides of the invention by recombinant techniques.

[0057] Host cells are genetically engineered (transduced or transformed or transfected) with the vectors of this invention which may be, for example, a cloning vector or an expression vector. The vector may be, for example, in the form of a plasmid, a viral particle, a phage, etc. The engineered host cells can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants or amplifying the hMLH1, hMLH2 and hMLH3 genes. The culture conditions, such as temperature, pH and the like, are those previously used with the host cell selected for expression, and will be apparent to the ordinarily skilled artisan.

[0058] The polynucleotides of the present invention may be employed for producing polypeptides by recombinant techniques. Thus, for example, the polynucleotide may be included in any one of a variety of expression vectors for expressing a polypeptide. Such vectors include chromosomal, nonchromosomal and synthetic DNA sequences, e.g., derivatives of SV40; bacterial plasmids; phage DNA; baculovirus; yeast plasmids; vectors derived from combinations of plasmids and phage DNA, viral DNA such as vaccinia,

adenovirus, fowl pox virus, and pseudorabies. However, any other vector may be used as long as it is replicable and viable in the host.

[0059] The appropriate DNA sequence may be inserted into the vector by a variety of procedures. In general, the DNA sequence is inserted into an appropriate restriction endonuclease site(s) by procedures known in the art. Such procedures and others are deemed to be within the scope of those skilled in the art.

[0060] The DNA sequence in the expression vector is operatively linked to an appropriate expression control sequence(s) (promoter) to direct mRNA synthesis. As representative examples of such promoters, there may be mentioned: LTR or SV40 promoter, the E. coli. lac or trp, the phage lambda P_L promoter and other promoters known to control expression of genes in prokaryotic or eukaryotic cells or their viruses. The expression vector also contains a ribosome binding site for translation initiation and a transcription terminator. The vector may also include appropriate sequences for amplifying expression.

[0061] In addition, the expression vectors preferably contain one or more selectable marker genes to provide a phenotypic trait for selection of transformed host cells such as dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, or such as tetracycline or ampicillin resistance in E. coli.

[0062] The vector containing the appropriate DNA sequence as hereinabove described, as well as an appropriate promoter or control sequence, may be employed to transform an appropriate host to permit the host to express the proteins.

[0063] As representative examples of appropriate hosts, there may be mentioned: bacterial cells, such as E. coli, Streptomyces, Salmonella typhimurium; fungal cells, such as yeast; insect cells such as Drosophila S2 and Spodoptera Sf9; animal cells such as CHO, COS or Bowes melanoma; adenoviruses; plant cells, etc. The selection of an appropriate host is deemed to be within the scope of those skilled in the art from the teachings herein.

[0064] More particularly, the present invention also includes recombinant constructs comprising one or more of the sequences as broadly described above. The constructs comprise a vector, such as a plasmid or viral vector, into which a sequence of the

invention has been inserted, in a forward or reverse orientation. In a preferred aspect of this embodiment, the construct further comprises regulatory sequences, including, for example, a promoter, operably linked to the sequence. Large numbers of suitable vectors and promoters are known to those of skill in the art, and are commercially available. The following vectors are provided by way of example. Bacterial: pQE70, pQE60, pQE-9 (Qiagen, Inc.), pbs, pD10, phagescript, psiX174, pbluescript SK, pbsks, pNH8A, pNH16a, pNH18A, pNH46A (Stratagene); ptrc99a, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia). Eukaryotic: pWLNEO, pSV2CAT, pOG44, pXT1, pSG (Stratagene) pSVK3, pBPV, pMSG, pSVL (Pharmacia). However, any other plasmid or vector may be used as long as they are replicable and viable in the host.

[0065] Promoter regions can be selected from any desired gene using CAT (chloramphenicol transferase) vectors or other vectors with selectable markers. Two appropriate vectors are pKK232-8 and pCM7. Particular named bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda P_R, P_L and TRP. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector and promoter is well within the level of ordinary skill in the art.

[0066] In a further embodiment, the present invention relates to host cells containing the above-described constructs. The host cell can be a higher eukaryotic cell, such as a mammalian cell, or a lower eukaryotic cell, such as a yeast cell, or the host cell can be a prokaryotic cell, such as a bacterial cell. Introduction of the construct into the host cell can be effected by calcium phosphate transfection, DEAE-Dextran mediated transfection, or electroporation (Davis, L., Dibner, M., Battey, I., Basic Methods in Molecular Biology, (1986)).

[0067] The constructs in host cells can be used in a conventional manner to produce the gene product encoded by the recombinant sequence. Alternatively, the polypeptides of the invention can be synthetically produced by conventional peptide synthesizers.

[0068] Mature proteins can be expressed in mammalian cells, yeast, bacteria, or other cells under the control of appropriate promoters. Cell-free translation systems can also be employed to produce such proteins using RNAs derived from the DNA constructs of the

present invention. Appropriate cloning and expression vectors for use with prokaryotic and eukaryotic hosts are described by Sambrook, et al., Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor, N.Y., (1989), the disclosure of which is hereby incorporated by reference.

[0069] Transcription of the DNA encoding the polypeptides of the present invention by higher eukaryotes is increased by inserting an enhancer sequence into the vector. Enhancers are cis-acting elements of DNA, usually about from 10 to 300 bp that act on a promoter to increase its transcription. Examples including the SV40 enhancer on the late side of the replication origin bp 100 to 270, a cytomegalovirus early promoter enhancer, the polyoma enhancer on the late side of the replication origin, and adenovirus enhancers.

[0070] Generally, recombinant expression vectors will include origins of replication and selectable markers permitting transformation of the host cell, e.g., the ampicillin resistance gene of E. coli and S. cerevisiae TRP1 gene, and a promoter derived from a highly-expressed gene to direct transcription of a downstream structural sequence. Such promoters can be derived from operons encoding glycolytic enzymes such as 3-phosphoglycerate kinase (PGK), -factor, acid phosphatase, or heat shock proteins, among others. The heterologous structural sequence is assembled in appropriate phase with translation initiation and termination sequences. Optionally, the heterologous sequence can encode a fusion protein including an N-terminal identification peptide imparting desired characteristics, e.g., stabilization or simplified purification of expressed recombinant product.

[0071] Useful expression vectors for bacterial use are constructed by inserting a structural DNA sequence encoding a desired protein together with suitable translation initiation and termination signals in operable reading phase with a functional promoter. The vector will comprise one or more phenotypic selectable markers and an origin of replication to ensure maintenance of the vector and to, if desirable, provide amplification within the host. Suitable prokaryotic hosts for transformation include E. coli, Bacillus subtilis, Salmonella typhimurium and various species within the genera Pseudomonas, Streptomyces, and Staphylococcus, although others may also be employed as a matter of choice.

[0072] As a representative but nonlimiting example, useful expression vectors for bacterial use can comprise a selectable marker and bacterial origin of replication derived from commercially available plasmids comprising genetic elements of the well known cloning vector pBR322 (ATCC 37017). Such commercial vectors include, for example, pKK223-3 (Pharmacia Fine Chemicals, Uppsala, Sweden) and GEM1 (Promega Biotec, Madison, WI, USA). These pBR322 "backbone" sections are combined with an appropriate promoter and the structural sequence to be expressed.

[0073] Following transformation of a suitable host strain and growth of the host strain to an appropriate cell density, the selected promoter is induced by appropriate means (e.g., temperature shift or chemical induction) and cells are cultured for an additional period.

[0074] Cells are typically harvested by centrifugation, disrupted by physical or chemical means, and the resulting crude extract retained for further purification.

[0075] Microbial cells employed in expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents, such methods are well know to those skilled in the art.

[0076] Various mammalian cell culture systems can also be employed to express recombinant protein. Examples of mammalian expression systems include the COS-7 lines of monkey kidney fibroblasts, described by Gluzman, Cell, 23:175 (1981), and other cell lines capable of expressing a compatible vector, for example, the C127, 3T3, CHO, HeLa and BHK cell lines. Mammalian expression vectors will comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. DNA sequences derived from the SV40 splice, and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

[0077] The polypeptides can be recovered and purified from recombinant cell cultures by methods including ammonium sulfate or ethanol precipitation, acid extraction, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity chromatography, hydroxylapatite chromatography and lectin chromatography. Protein refolding steps can be used, as necessary, in completing

configuration of the mature protein. Finally, high performance liquid chromatography (HPLC) can be employed for final purification steps.

[0078] The polypeptides of the present invention may be a naturally purified product, or a product of chemical synthetic procedures, or produced by recombinant techniques from a prokaryotic or eukaryotic host (for example, by bacterial, yeast, higher plant, insect and mammalian cells in culture). Depending upon the host employed in a recombinant production procedure, the polypeptides of the present invention may be glycosylated or may be non-glycosylated.

[0079] In accordance with a further aspect of the invention, there is provided a process for determining susceptibility to cancer, in particular, a hereditary cancer. Thus, a mutation in a human repair protein, which is a human homolog of *mutL*, and in particular those described herein, indicates a susceptibility to cancer, and the nucleic acid sequences encoding such human homologs may be employed in an assay for ascertaining such susceptibility. Thus, for example, the assay may be employed to determine a mutation in a human DNA repair protein as herein described, such as a deletion, truncation, insertion, frame shift, etc., with such mutation being indicative of a susceptibility to cancer.

[0080] A mutation may be ascertained for example, by a DNA sequencing assay. Tissue samples, including but not limited to blood samples are obtained from a human patient. The samples are processed by methods known in the art to capture the RNA. First strand cDNA is synthesized from the RNA samples by adding an oligonucleotide primer consisting of polythymidine residues which hybridize to the polyadenosine stretch present on the mRNA's. Reverse transcriptase and deoxynucleotides are added to allow synthesis of the first strand cDNA. Primer sequences are synthesized based on the DNA sequence of the DNA repair protein of the invention. The primer sequence is generally comprised of 15 to 30 and preferably from 18 to 25 consecutive bases of the human DNA repair gene. Table 1 sets forth an illustrative example of oligonucleotide primer sequences based on hMLH1. The primers are used in pairs (one "sense" strand and one "anti-sense") to amplify the cDNA from the patients by the PCR method (Saiki *et al.*, Nature, 324:163-166 (1986)) such that three overlapping fragments of the patient's cDNA's for such protein are generated. Table 1 also shows a list of preferred primer sequence pairs. The overlapping

fragments are then subjected to dideoxynucleotide sequencing using a set of primer sequences synthesized to correspond to the base pairs of the cDNA's at a point approximately every 200 base pairs throughout the gene.

TABLE 1

Primer Sequences used to amplify gene region using PCR

Name	SEQ ID NO:	Start Site and Arrangement	Sequence
758	7	sense-(-41)*	GTTGAACATCTAGACGTCTC
1319	8	sense-8	TCGTGGCAGGGGTTATTCG
1321	9	sense-619	CTACCCAATGCCTCAACCG
1322	10	sense-677	GAGAACTGATAGAAATTGGATG
1314	11	sense-1548	GGGACATGAGGTTCTCCG
1323	12	sense-1593	GGGCTGTGTGAATCCTCAG
773	13	anti-53	CGGTTCACCACTGTCTCGTC
1313	14	anti-971	TCCAGGATGCTCTCCTCG
1320	15	anti-1057	CAAGTCCTGGTAGCAAAGTC
1315	16	anti-1760	ATGGCAAGGTCAAAGAGCG
1316	17	anti-1837	CAACAATGTATTCAGXAAGTCC
1317	18	anti-2340	TTGATACAACACTTTGTATCG
1318	19	anti-2415	GGAATACTATCAGAAGGCAAG

[0081] * Numbers corresponding to location along nucleotide sequence of Figure 1 where ATG is number 1.

[0082] Preferred primer sequences pairs:

758, 1313
1319, 1320
660, 1909
725, 1995
1680, 2536
1727, 2610

[0083] The nucleotide sequences shown in Table 1 represent SEQ ID NO:7 through 19, respectively.

[0084] Table 2 lists representative examples of oligonucleotide primer sequences (sense and anti-sense) which may be used, and preferably the entire set of primer sequences are used for sequencing to determine where a mutation in the patient DNA repair protein may be. The primer sequences may be from 15 to 30 bases in length and are

preferably between 18 and 25 bases in length. The sequence information determined from the patient is then compared to non-mutated sequences to determine if any mutations are present.

TABLE 2

Primer Sequences Used to Sequence the Amplified Fragments

Name	SEQ ID NO:	Start Site and Arrangement	Sequence
5282	~~seq01~~ 20	sense-377*	ACAGAGCAAGTTACTCAGATG
5283	~~seq02~~ 21	sense-552	GTACACAATGCAGGCATTAG
5284	~~seq03~~ 22	sense-904	AATGTGGATGTTAATGTGCAC
5285	~~seq04~~ 23	sense-1096	CTGACCTCGTCTTCCTAC
5286	~~seq05~~ 24	sense-1276	CAGCAAGATGAGGAGATGC
5287	~~seq06~~ 25	sense-1437	GGAAATGGTGGAAGATGATTC
5288	~~seq07~~ 26	sense-1645	CTTCTCAACACCAAGC
5289	~~seq08~~ 27	sense-1895	GAAATTGATGAGGAAGGGAAC
5295	~~seq09~~ 28	sense-1921	CTTCTGATTGACAACTATGTGC
5294	~~seq10~~ 29	sense-2202	CACAGAAGATGGAAATATCCTG
293	~~seq11~~ 30	sense-2370	GTGTTGGTAGCACTTAAGAC
5291	~~seq12~~ 31	anti-525	TTTCCCATATTCTTCACTTG
5290	~~seq13~~ 32	anti-341	GTAACATGAGCCACATGGC
5292	~~seq14~~ 33	anti-46	CCACTGTCTCGTCCAGCCG

[0085] * Numbers corresponding to location along nucleotide sequence of Figure 1 where ATG is number 1.

[0086] The nucleotide sequences shown in Table 2 represent SEQ ID NO:20 through 33, respectively.

[0087] In another embodiment, the primer sequences from Table 2 could be used in the PCR method to amplify a mutated region. The region could be sequenced and used as a diagnostic to predict a predisposition to such mutated genes.

[0088] Alternatively, the assay to detect mutations in the genes of the present invention may be performed by genetic testing based on DNA sequence differences achieved by

detection of alteration in electrophoretic mobility of DNA fragments in gels with or without denaturing agents. Small sequence deletions and insertions can be visualized by high resolution gel electrophoresis. DNA fragments of different sequences may be distinguished on denaturing formamide gradient gels in which the mobilities of different DNA fragments are retarded in the gel at different positions according to their specific melting or partial melting temperatures (see, e.g., Myers *et al.*, Science, 230:1242 (1985)).

[0089] Sequence changes at specific locations may also be revealed by nuclease protection assays, such as RNase and S1 protection or the chemical cleavage method (e.g., Cotton *et al.*, PNAS, USA, 85:4397-4401 (1985)). Perfectly matched sequences can be distinguished from mismatched duplexes by RNase A digestion or by differences in melting temperatures.

[0090] Thus, the detection of a specific DNA sequence may be achieved by methods such as hybridization, RNase protection, chemical cleavage, Western Blot analysis, direct DNA sequencing or the use of restriction enzymes, (e.g., Restriction Fragment Length Polymorphisms (RFLP)) and Southern blotting of genomic DNA.

[0091] In addition to more conventional gel-electrophoresis and DNA sequencing, mutations can also be detected by *in situ* analysis.

[0092] The polypeptides may also be employed to treat cancers or to prevent cancers, by expression of such polypeptides *in vivo*, which is often referred to as "gene therapy."

[0093] Thus, for example, cells from a patient may be engineered with a polynucleotide (DNA or RNA) encoding a polypeptide *ex vivo*, with the engineered cells then being provided to a patient to be treated with the polypeptide. Such methods are well-known in the art. For example, cells may be engineered by procedures known in the art by use of a retroviral particle containing RNA encoding a polypeptide of the present invention.

[0094] Similarly, cells may be engineered *in vivo* for expression of a polypeptide *in vivo* by, for example, procedures known in the art. As known in the art, a producer cell for producing a retroviral particle containing RNA encoding the polypeptide of the present invention may be administered to a patient for engineering cells *in vivo* and expression of the polypeptide *in vivo*. These and other methods for administering a polypeptide of the

present invention by such method should be apparent to those skilled in the art from the teachings of the present invention. For example, the expression vehicle for engineering cells may be other than a retrovirus, for example, an adenovirus which may be used to engineer cells in vivo after combination with a suitable delivery vehicle.

[0095] Retroviruses from which the retroviral plasmid vectors hereinabove mentioned may be derived include, but are not limited to, Moloney Murine Leukemia Virus, spleen necrosis virus, retroviruses such as Rous Sarcoma Virus, Harvey Sarcoma Virus, avian leukosis virus, gibbon ape leukemia virus, human immunodeficiency virus, adenovirus, Myeloproliferative Sarcoma Virus, and mammary tumor virus. In one embodiment, the retroviral plasmid vector is derived from Moloney Murine Leukemia Virus.

[0096] The vector includes one or more promoters. Suitable promoters which may be employed include, but are not limited to, the retroviral LTR; the SV40 promoter; and the human cytomegalovirus (CMV) promoter described in Miller, et al., Biotechniques, Vol. 7, No. 9, 980-990 (1989), or any other promoter (e.g., cellular promoters such as eukaryotic cellular promoters including, but not limited to, the histone, pol III, and -actin promoters). Other viral promoters which may be employed include, but are not limited to, adenovirus promoters, thymidine kinase (TK) promoters, and B19 parvovirus promoters. The selection of a suitable promoter will be apparent to those skilled in the art from the teachings contained herein.

[0097] The nucleic acid sequence encoding the polypeptide of the present invention is under the control of a suitable promoter. Suitable promoters which may be employed include, but are not limited to, adenoviral promoters, such as the adenoviral major late promoter; or hetorologous promoters, such as the cytomegalovirus (CMV) promoter; the respiratory syncytial virus (RSV) promoter; inducible promoters, such as the MMT promoter, the metallothionein promoter; heat shock promoters; the albumin promoter; the ApoAI promoter; human globin promoters; viral thymidine kinase promoters, such as the Herpes Simplex thymidine kinase promoter; retroviral LTRs (including the modified retroviral LTRs hereinabove described); the -actin promoter; and human growth hormone promoters. The promoter also may be the native promoter which controls the gene encoding the polypeptide.

[0098] The retroviral plasmid vector is employed to transduce packaging cell lines to form producer cell lines. Examples of packaging cells which may be transfected include, but are not limited to, the PE501, PA317, -2, -AM, PA12, T19-14X, VT-19-17-H2, CRE, CRIP, GP+E-86, GP+envAm12, and DAN cell lines as described in Miller, Human Gene Therapy, Vol. 1, pgs. 5-14 (1990), which is incorporated herein by reference in its entirety. The vector may transduce the packaging cells through any means known in the art. Such means include, but are not limited to, electroporation, the use of liposomes, and $CaPO_4$ precipitation. In one alternative, the retroviral plasmid vector may be encapsulated into a liposome, or coupled to a lipid, and then administered to a host.

[0099] The producer cell line generates infectious retroviral vector particles which include the nucleic acid sequence(s) encoding the polypeptides. Such retroviral vector particles then may be employed, to transduce eukaryotic cells, either *in vitro* or *in vivo*. The transduced eukaryotic cells will express the nucleic acid sequence(s) encoding the polypeptide. Eukaryotic cells which may be transduced include, but are not limited to, embryonic stem cells, embryonic carcinoma cells, as well as hematopoietic stem cells, hepatocytes, fibroblasts, myoblasts, keratinocytes, endothelial cells, and bronchial epithelial cells.

[0100] Each of the cDNA sequences identified herein or a portion thereof can be used in numerous ways as polynucleotide reagents. The sequences can be used as diagnostic probes for the presence of a specific mRNA in a particular cell type. In addition, these sequences can be used as diagnostic probes suitable for use in genetic linkage analysis (polymorphisms).

[0101] The sequences of the present invention are also valuable for chromosome identification. The sequence is specifically targeted to and can hybridize with a particular location on an individual human chromosome. Moreover, there is a current need for identifying particular sites on the chromosome. Few chromosome marking reagents based on actual sequence data (repeat polymorphisms) are presently available for marking chromosomal location. The mapping of DNAs to chromosomes according to the present invention is an important first step in correlating those sequences with genes associated with disease.

[0102] Briefly, sequences can be mapped to chromosomes by preparing PCR primers (preferably 15-25 bp) from the cDNA. Computer analysis of the 3' untranslated region is used to rapidly select primers that do not span more than one exon in the genomic DNA, thus complicating the amplification process. These primers are then used for PCR screening of somatic cell hybrids containing individual human chromosomes. Only those hybrids containing the human gene corresponding to the primer will yield an amplified fragment.

[0103] PCR mapping of somatic cell hybrids is a rapid procedure for assigning a particular DNA to a particular chromosome. Using the present invention with the same oligonucleotide primers, sublocalization can be achieved with panels of fragments from specific chromosomes or pools of large genomic clones in an analogous manner. Other mapping strategies that can similarly be used to map to its chromosome include *in situ* hybridization, prescreening with labeled flow-sorted chromosomes and preselection by hybridization to construct chromosome-specific cDNA libraries.

[0104] Fluorescence *in situ* hybridization (FISH) of a cDNA clone to a metaphase chromosomal spread can be used to provide a precise chromosomal location in one step. This technique can be used with cDNA as short as 50 or 60 bases. For a review of this technique, see Verma et al., Human Chromosomes: a Manual of Basic Techniques, Pergamon Press, New York (1988).

[0105] Once a sequence has been mapped to a precise chromosomal location, the physical position of the sequence on the chromosome can be correlated with genetic map data. Such data are found, for example, in V. McKusick, Mendelian Inheritance in Man (available on line through Johns Hopkins University Welch Medical Library). The relationship between genes and diseases that have been mapped to the same chromosomal region are then identified through linkage analysis (coinheritance of physically adjacent genes).

[0106] Next, it is necessary to determine the differences in the cDNA or genomic sequence between affected and unaffected individuals. If a mutation is observed in some or all of the affected individuals but not in any normal individuals, then the mutation is likely to be the causative agent of the disease.

[0107] With current resolution of physical mapping and genetic mapping techniques, a cDNA precisely localized to a chromosomal region associated with the disease could be one of between 50 and 500 potential causative genes. (This assumes 1 megabase mapping resolution and one gene per 20 kb).

[0108] hMLH2 has been localized using a genomic P1 clone (1670) which contained the 5' region of the hMLH2 gene. Detailed analysis of human metaphase chromosome spreads, counterstained to reveal banding, indicated that the hMLH2 gene was located within bands 2q32. Likewise, hMLH3 was localized using a genomic P1 clone (2053) which contained the 3' region of the hMLH3 gene. Detailed analysis of human metaphase chromosome spreads, counterstained to reveal banding, indicated that the hMLH3 gene was located within band 7p22, the most distal band on chromosome 7. Analysis with a variety of genomic clones showed that hMLH3 was a member of a subfamily of related genes, all on chromosome 7.

[0109] The polypeptides, their fragments or other derivatives, or analogs thereof, or cells expressing them can be used as an immunogen to produce antibodies thereto. These antibodies can be, for example, polyclonal or monoclonal antibodies. The present invention also includes chimeric, single chain, and humanized antibodies, as well as Fab fragments, or the product of an Fab expression library. Various procedures known in the art may be used for the production of such antibodies and fragments.

[0110] Antibodies generated against the polypeptides corresponding to a sequence of the present invention can be obtained by direct injection of the polypeptides into an animal or by administering the polypeptides to an animal, preferably a nonhuman. The antibody so obtained will then bind the polypeptides itself. In this manner, even a sequence encoding only a fragment of the polypeptides can be used to generate antibodies binding the whole native polypeptides. Such antibodies can then be used to isolate the polypeptide from tissue expressing that polypeptide.

[0111] For preparation of monoclonal antibodies, any technique which provides antibodies produced by continuous cell line cultures can be used. Examples include the hybridoma technique (Kohler and Milstein, 1975, Nature, 256:495-497), the trioma technique, the human B-cell hybridoma technique (Kozbor et al., 1983, Immunology

Today 4:72), and the EBV-hybridoma technique to produce human monoclonal antibodies (Cole, et al., 1985, in Monoclonal Antibodies and Cancer Therapy, Alan R. Liss, Inc., pp. 77-96).

[0112] Techniques described for the production of single chain antibodies (U.S. Patent 4,946,778) can be adapted to produce single chain antibodies to immunogenic polypeptide products of this invention. Also, transgenic mice may be used to express humanized antibodies to immunogenic polypeptide products of this invention.

[0113] The present invention will be further described with reference to the following examples; however, it is to be understood that the present invention is not limited to such examples. All parts or amounts, unless otherwise specified, are by weight.

[0114] In order to facilitate understanding of the following examples certain frequently occurring methods and/or terms will be described.

[0115] "Plasmids" are designated by a lower case p preceded and/or followed by capital letters and/or numbers. The starting plasmids herein are either commercially available, publicly available on an unrestricted basis, or can be constructed from available plasmids in accord with published procedures. In addition, equivalent plasmids to those described are known in the art and will be apparent to the ordinarily skilled artisan.

[0116] "Digestion" of DNA refers to catalytic cleavage of the DNA with a restriction enzyme that acts only at certain sequences in the DNA. The various restriction enzymes used herein are commercially available and their reaction conditions, cofactors and other requirements were used as would be known to the ordinarily skilled artisan. For analytical purposes, typically 1 μg of plasmid or DNA fragment is used with about 2 units of enzyme in about 20 μl of buffer solution. For the purpose of isolating DNA fragments for plasmid construction, typically 5 to 50 μg of DNA are digested with 20 to 250 units of enzyme in a larger volume. Appropriate buffers and substrate amounts for particular restriction enzymes are specified by the manufacturer. Incubation times of about 1 hour at 37°C are ordinarily used, but may vary in accordance with the supplier's instructions. After digestion the reaction is electrophoresed directly on a polyacrylamide gel to isolate the desired fragment.

[0117] Size separation of the cleaved fragments is performed using 8 percent polyacrylamide gel described by Goeddel, D. et al., Nucleic Acids Res., 8:4057 (1980).

[0118] "Oligonucleotides" refers to either a single stranded polydeoxynucleotide or two complementary polydeoxynucleotide strands which may be chemically synthesized. Such synthetic oligonucleotides have no 5' phosphate and thus will not ligate to another oligonucleotide without adding a phosphate with an ATP in the presence of a kinase. A synthetic oligonucleotide will ligate to a fragment that has not been dephosphorylated.

[0119] "Ligation" refers to the process of forming phosphodiester bonds between two double stranded nucleic acid fragments (Maniatis, T., et al., Id., p. 146). Unless otherwise provided, ligation may be accomplished using known buffers and conditions with 10 units to T4 DNA ligase ("ligase") per 0.5 μg of approximately equimolar amounts of the DNA fragments to be ligated.

[0120] Unless otherwise stated, transformation was performed as described in the method of Graham, F. and Van der Eb, A., Virology, 52:456-457 (1973).

Examples

Example 1: Bacterial Expression of hMLH1

[0121] The full length DNA sequence encoding human DNA mismatch repair protein hMLH1, ATCC # 75649, is initially amplified using PCR oligonucleotide primers corresponding to the 5' and 3' ends of the DNA sequence to synthesize insertion fragments. The 5' oligonucleotide primer has the sequence 5' CGGGATCCATGTCGTTCGTGGCAGGG 3' (SEQ ID NO:34), contains a BamHI restriction enzyme site followed by 18 nucleotides of hMLH1 coding sequence following the initiation codon; the 3' sequence 5' GCTCTAGATTAACACCTCTCAAAGAC 3' (SEQ ID NO:35) contains complementary sequences to an XbaI site and is at the end of the gene. The restriction enzyme sites correspond to the restriction enzyme sites on the bacterial expression vector pQE-9. (Qiagen, Inc., Chatsworth, CA). The plasmid vector encodes antibiotic resistance (Ampr), a bacterial origin of replication (ori), an IPTG-regulatable promoter/operator (P/O), a ribosome binding site (RBS), a 6-histidine tag (6-His) and restriction enzyme cloning sites. The pQE-9 vector is digested with BamHI and

XbaI and the insertion fragments are then ligated into the pQE-9 vector maintaining the reading frame initiated at the bacterial RBS. The ligation mixture is then used to transform the E. coli strain M15/rep4 (Qiagen, Inc.) which contains multiple copies of the plasmid pREP4, which expresses the lacI repressor and also confers kanamycin resistance (Kanr). Transformants are identified by their ability to grow on LB plates and ampicillin/kanamycin resistant colonies are selected. Plasmid DNA is isolated and confirmed by restriction analysis. Clones containing the desired constructs are grown overnight (O/N) in liquid culture in LB media supplemented with both Amp (100 ug/ml) and Kan (25 ug/ml). Tho O/N culture is used to inoculate a large culture at a ratio of 1:100 to 1:250. The cells are grown to an optical density 600 (O.D.600) of between 0.4 and 0.6. IPTG (Isopropyl-B-D-thiogalacto pyranoside) is then added to a final concentration of 1 mM. IPTG induces by inactivating the lacI repressor, clearing the P/O leading to increased gene expression. Cells are grown an extra 3 to 4 hours. Cells are then harvested by centrifugation (20 mins at 6000Xg). The cell pellet is solubilized in the chaotropic agent 6 Molar Guanidine HCl. After clarification, solubilized hMLH1 is purified from this solution by chromatography on a Nickel-Chelate column under conditions that allow for tight binding by proteins containing the 6-His tag (Hochuli, E. et al., Genetic Engineering, Principles & Methods, 12:87-98 (1990). Protein renaturation out of GnHCl can be accomplished by several protocols (Jaenicke, R. and Rudolph, R., Protein Structure - A Practical Approach, IRL Press, New York (1990)). Initially, step dialysis is utilized to remove the GnHCL. Alternatively, the purified protein isolated from the Ni-chelate column can be bound to a second column over which a decreasing linear GnHCL gradient is run. The protein is allowed to renature while bound to the column and is subsequently eluted with a buffer containing 250 mM Imidazole, 150 mM NaCl, 25 mM Tris-HCl pH 7.5 and 10% Glycerol. Finally, soluble protein is dialyzed against a storage buffer containing 5 mM Ammonium Bicarbonate. The purified protein was analyzed by SDS-PAGE.

Example 2: Spontaneous Mutation Assay for Detection of the Expression of hMLH1, hMLH2 and hMLH3 and Complementation to the E.coli mutl

[0122] The pQE9hMLH1, pQE9hMLH2 or pQE9hMLH3/GW3733, transformants were subjected to the spontaneous mutation assay. The plasmid vector pQE9 was also transformed to AB1157 (k-12, *argE3 hisG4,LeuB6 proA2 thr-1 ara-1 rpsL31 supE44 tsx-33*) and GW3733 to use as the positive and negative control respectively.

[0123] Fifteen 2 ml cultures, inoculated with approximately 100 to 1000 *E. coli,* were grown $2x10^8$ cells per ml in LB ampicillin medium at 37°C. Ten microliters of each culture were diluted and plated on the LB ampicillin plates to measure the number of viable cells. The rest of the cells from each culture were then concentrated in saline and plated on minimal plates lacking of arginine to measure reversion of Arg^+. In Table 3, the mean number of mutations per culture (m) was calculated from the median number (r) of mutants per distribution, according to the equation $(r/m)-\ln(m) = 1.24$ (Lea et al., J. Genetics 49:264-285 (1949)). Mutation rates per generation were recorded as m/N, with N representing the average number of cells per culture.

TABLE 3

Spontaneous Mutation Rates

Strain	Mutation/generation
AB1157+vector	(5.6±0.1) x 10-9a
GW3733+vector	(1.1±0.2) x 10-6a
GW3733+phMLH1	(3.7±1.3 x 10-7a
GW3733+phMLH2	(3.1±0.6) x 10-7b
GW3733+phMLH3	(2.1±0.8) x 10-7b

a: Average of three experiments.

b: Average of four experiments.

[0124] The functional complementation result showed that the human *mutL* can partially rescue the E.coli *mutL* mutator phenotype, suggesting that the human *mutL* is not only successfully expressed in a bacterial expression system, but also functions in bacteria.

Example 3: Chromosomal Mapping of the hMLH1

[0125] An oligonucleotide primer set was designed according to the sequence at the 5' end of the cDNA for HMLH1. This primer set would span a 94 bp segment. This primer set was used in a polymerase chain reaction under the following set of conditions :

30 seconds, 95 degrees C

1 minute, 56 degrees C

1 minute, 70 degrees C

This cycle was repeated 32 times followed by one 5 minute cycle at 70 degrees C. Human, mouse, and hamster DNA were used as template in addition to a somatic cell hybrid panel (Bios, Inc). The reactions were analyzed on either 8% polyacrylamide gels or 3.5 %

agarose gels. A 94 base pair band was observed in the human genomic DNA sample and in the somatic cell hybrid sample corresponding to chromosome 3. In addition, using various other somatic cell hybrid genomic DNA, the hMLH1 gene was localized to chromosome 3p.

Example 4: Method for Determination of mutation of hMLH1 gene in HNPCC kindred

[0126] cDNA was produced from RNA obtained from tissue samples from persons who are HNPCC kindred and the cDNA was used as a template for PCR, employing the primers 5' GCATCTAGACGTTTCCTTGGC 3' (SEQ ID NO:36) and 5' CATCCAAGCTTCTGTTCCCG 3' (SEQ ID NO:37), allowing amplification of codons 1 to 394 of Figure 1; 5' GGGGTGCAGCAGCACATCG 3' (SEQ ID NO:38) and 5' GGAGGCAGAATGTGTGAGCG 3' (SEQ ID NO:39), allowing amplification of codons 326 to 729 of Figure 1 (SEQ ID NO:2); and 5' TCCCAAAGAAGGACTTGCT 3' (SEQ ID NO:40) and 5' AGTATAAGTCTTAAGTGCTACC 3' (SEQ ID NO:41), allowing amplification of codons 602 to 756 plus 128 nt of 3'- untranslated sequences of Figure 1 (SEQ ID NO:2). The PCR conditions for all analyses used consisted of 35 cycles at 95 C for 30 seconds, 52-58 C for 60 to 120 seconds, and 70 C for 60 to 120 seconds, in the buffer solution described in San Sidransky, D. *et al.*, Science, 252:706 (1991). PCR products were sequenced using primers labeled at their 5' end with T4 polynucleotide kinase, employing SequiTherm Polymerase (Epicentre Technologies). The intron-exon borders of selected exons were also determined and genomic PCR products analyzed to confirm the results. PCR products harboring suspected mutations were then cloned and sequenced to validate the results of the direct sequencing. PCR products were cloned into T-tailed vectors as described in Holton, T.A. and Graham, M.W., Nucleic Acids Research, 19:1156 (1991) and sequenced with T7 polymerase (United States Biochemical). Affected individuals from seven kindreds all exhibited a heterozygous deletion of codons 578 to 632 of the hMLH1 gene. The derivation of five of these seven kindreds could be traced to a common ancestor. The genomic sequences surrounding codons 578-632 were determined by cycle-sequencing of the P1 clones (a human genomic P1 library which

contains the entire hMLH1 gene (Genome Systems)) using SequiTherm Polymerase, as described by the manufacturer, with the primers were labeled with T4 polynucleotide kinase, and by sequencing PCR products of genomic DNA. The primers used to amplify the exon containing codons 578-632 were 5' TTTATGGTTTCTCACCTGCC 3' (SEQ ID NO:42) and 5' GTTATCTGCCCACCTCAGC 3' (SEQ ID NO:43). The PCR product included 105 bp of intron C sequence upstream of the exon and 117 bp downstream. No mutations in the PCR product were observed in the kindreds, so the deletion in the RNA was not due to a simple splice site mutation. Codons 578 to 632 were found to constitute a single exon which was deleted from the gene product in the kindreds described above. This exon contains several highly conserved amino acids.

[0127] In a second family (L7), PCR was performed using the above primers and a 4bp deletion was observed beginning at the first nucleotide (nt) of codon 727. This produced a frame shift with a new stop codon 166 nt downstream, resulting in a substitution of the carboxy-terminal 29 amino acids of hMLH1 with 53 different amino acids, some encoded by nt normally in the 3' untranslated region.

[0128] A different mutation was found in a different kindred (L2516) after PCR using the above primers, the mutation consisting of a 4bp insert between codons 755 and 756. This insertion resulted in a frame shift and extension of the ORF to include 102 nucleotides (34 amino acids) downstream of the normal termination codon. The mutations in both kindreds L7 and L2516 were therefore predicted to alter the C-terminus of hMLH1.

[0129] A possible mutation in the hMLH1 gene was determined from alterations in size of the encoded protein, where kindreds were too few for linkage studies. The primers used for coupled transcription-translation of hMLH1 were 5' GGATCCTAATACGACTCACTATAGGGAGACCACCATGGCATCTAGACGTTT CCCTTGGC 3' (SEQ ID NO:44) and 5' CATCCAAGCTTCTGTTCCCG 3' (SEQ ID NO:45) for codons 1 to 394 of Figure 1 and 5' GGATCCTAATACGACTCACTATAGGGAGACCACCATGGGGGTGCAGCAGCA CATCG 3' (SEQ ID NO:46) and 5' GGAGGCAGAATGTGTGAGCG 3' (SEQ ID NO:47) for codons 326 to 729 of Figure 1 (SEQ ID NO:2). The resultant PCR products had signals for transcription by T7 RNA polymerase and for the initiation of translation at their

5' ends. RNA from lymphoblastoid cells of patients from 18 kindreds was used to amplify two products, extending from codon 1 to codon 394 or from codon 326 to codon 729, respectively. The PCR products were then transcribed and translated in vitro, making use of transcription-translation signals incorporated into the PCR primers. PCR products were used as templates in coupled transcription-translation reactions performed as described by Powell, S.M. et al., New England Journal of Medicine, 329:1982, (1993), using 40 micro CI of 35S labeled methionine. Samples were diluted in sample buffer, boiled for five minutes and analyzed by electropheresis on sodium dodecyl sulfate-polyacrylamide gels containing a gradient of 10% to 20% acrylamide. The gels were dried and subjected to radiography. All samples exhibited a polypeptide of the expected size, but an abnormally migrating polypeptide was additionally found in one case. The sequence of the relevant PCR product was determined and found to include a 371 bp deletion beginning at the first nucleotide (nt) of codon 347. This alteration was present in heterozygous form, and resulted in a frame shift in a new stop codon 30 nt downstream of codon 346, thus explaining the truncated polypeptide observed.

[0130] Four colorectal tumor cell lines manifesting microsatellite instability were examined. One of the four (cell line H6) showed no normal peptide in this assay and produced only a short product migrating at 27 kd. The sequence of the corresponding cDNA was determined and found to harbor a C to A transversion at codon 252, resulting in the substitution of a termination codon for serine. In accord with the translational analyses, no band at the normal C position was identified in the cDNA or genomic DNA from this tumor, indicating that it was devoid of a functional hMLH1 gene.

[0131] Table 4 sets forth the results of these sequencing assays. Deletions were found in those people who were known to have a family history of the colorectal cancer. More particularly, 9 of 10 families showed an hMLH1 mutation.

TABLE 4

Summary of Mutations in *hMLH1*

Sample	Codon	cDNA Nucleotide Change	Predicted Coding Change
Kindreds F2, F3, F6, F8, F10, F11, F52	578-632	165 bp deletion	In-frame deletion
Kindred L7	727/728	4 bp deletion (TCACACATTC to TCATTCT)	Frameshift and substitution of new amino acids
Kindred L2516	755/756	4 bp insertion (GTGTTAA to GTGTTTGTTAA)	Extension of C-terminus
Kindred RA	347	371 bp deletion	Frameshift/ Truncation
H6 Colorectal Tumor	252	Transversion (TCA to TAA)	Serine to Stop

Example 5: Bacterial Expression and Purification of hMLH2

[0132] The DNA sequence encoding hMLH2, ATCC #75651, is initially amplified using PCR oligonucleotide primers corresponding to the 5' and 3' ends of the DNA sequence to synthesize insertion fragments. The 5' oligonucleotide primer has the sequence 5' CGGGATCCATGAAACAATTGCCTGCGGC 3' (SEQ ID NO:48) contains a BamHI restriction enzyme site followed by 17 nucleotides of hMLH2 following the initiation codon. The 3' sequence 5' GCTCTAGACCAGACTCATGCTGTTTT 3' (SEQ ID NO:49) contains complementary sequences to an XbaI site and is followed by 18

nucleotides of hMLH2. The restriction enzyme sites correspond to the restriction enzyme sites on the bacterial expression vector pQE-9 (Qiagen, Inc. Chatsworth, CA). pQE-9 encodes antibiotic resistance (Amp^r), a bacterial origin of replication (ori), an IPTG-regulatable promoter operator (P/O), a ribosome binding site (RBS), a 6-His tag and restriction enzyme sites. The amplified sequences and pQE-9 are then digested with BamHI and XbaI. The amplified sequences are ligated into pQE-9 and are inserted in frame with the sequence encoding for the histidine tag and the RBS. The ligation mixture is then used to transform E. coli strain M15/rep4 (Qiagen, Inc.) which contains multiple copies of the plasmid pREP4, which expresses the lacI repressor and also confers kanamycin resistance (Kan^r). Transformants are identified by their ability to grow on LB plates and ampicillin/kanamycin resistant colonies are selected. Plasmid DNA is isolated and confirmed by restriction analysis. Clones containing the desired constructs are grown overnight (O/N) in liquid culture in LB media supplemented with both Amp (100 ug/ml) and Kan (25 ug/ml). Tho O/N culture is used to inoculate a large culture at a ratio of 1:100 to 1:250. The cells are grown to an optical density 600 ($O.D._{600}$) of between 0.4 and 0.6. IPTG (Isopropyl-B-D-thiogalacto pyranoside) is then added to a final concentration of 1 mM. IPTG induces by inactivating the lacI repressor, clearing the P/O leading to increased gene expression. Cells are grown an extra 3 to 4 hours. Cells are then harvested by centrifugation (20 mins at 6000Xg). The cell pellet is solubilized in the chaotropic agent 6 Molar Guanidine HCl. After clarification, solubilized hMLH2 is purified from this solution by chromatography on a Nickel-Chelate column under conditions that allow for tight binding by proteins containing the 6-His tag (Hochuli, E. et al., Genetic Engineering, Principles & Methods, 12:87-98 (1990). Protein renaturation out of GnHCl can be accomplished by several protocols (Jaenicke, R. and Rudolph, R., Protein Structure - A Practical Approach, IRL Press, New York (1990)). Initially, step dialysis is utilized to remove the GnHCL. Alternatively, the purified protein isolated from the Ni-chelate column can be bound to a second column over which a decreasing linear GnHCL gradient is run. The protein is allowed to renature while bound to the column and is subsequently eluted with a buffer containing 250 mM Imidazole, 150 mM NaCl, 25 mM Tris-HCl pH 7.5 and 10% Glycerol. Finally, soluble protein is dialyzed against a storage buffer

containing 5 mM Ammonium Bicarbonate. The purified protein was analyzed by SDS-PAGE.

Example 6: Bacterial Expression and Purification of hMLH3

[0133] The DNA sequence encoding hMLH3, ATCC #75650, is initially amplified using PCR oligonucleotide primers corresponding to the 5' and 3' ends of the DNA sequence to synthesize insertion fragments. The 5' oligonucleotide primer has the sequence 5' CGGGATCCATGGAGCGAGCTGAGAGC 3' (SEQ ID NO:50) contains a BamHI restriction enzyme site followed by 18 nucleotides of hMLH3 coding sequence starting from the presumed terminal amino acid of the processed protein. The 3' sequence 5' GCTCTAGAGTGAAG

[0134] ACTCTGTCT 3' (SEQ ID NO:51) contains complementary sequences to an XbaI site and is followed by 18 nucleotides of hMLH3. The restriction enzyme sites correspond to the restriction enzyme sites on the bacterial expression vector pQE-9 (Qiagen, Inc. Chatsworth, CA). pQE-9 encodes antibiotic resistance (Amp^r), a bacterial origin of replication (ori), an IPTG-regulatable promoter operator (P/O), a ribosome binding site (RBS), a 6-His tag and restriction enzyme sites. The amplified sequences and pQE-9 are then digested with BamHI and XbaI. The amplified sequences are ligated into pQE-9 and are inserted in frame with the sequence encoding for the histidine tag and the RBS. The ligation mixture was then used to transform E. coli strain M15/rep4 (Qiagen, Inc.) which contains multiple copies of the plasmid pREP4, which expresses the lacI repressor and also confers kanamycin resistance (Kan^r). Transformants are identified by their ability to grow on LB plates and ampicillin/kanamycin resistant colonies are selected. Plasmid DNA is isolated and confirmed by restriction analysis. Clones containing the desired constructs are grown overnight (O/N) in liquid culture in LB media supplemented with both Amp (100 ug/ml) and Kan (25 ug/ml). Tho O/N culture is used to inoculate a large culture at a ratio of 1:100 to 1:250. The cells are grown to an optical density 600 ($O.D._{600}$) of between 0.4 and 0.6. IPTG (Isopropyl-B-D-thiogalacto pyranoside) is then added to a final concentration of 1 mM. IPTG induces by inactivating the lacI repressor, clearing the P/O leading to increased gene expression. Cells are grown an extra 3 to 4

hours. Cells are then harvested by centrifugation (20 mins at 6000Xg). The cell pellet is solubilized in the chaotropic agent 6 Molar Guanidine HCl. After clarification, solubilized stanniocalcin is purified from this solution by chromatography on a Nickel-Chelate column under conditions that allow for tight binding by proteins containing the 6-His tag (Hochuli, E. et al., Genetic Engineering, Principles & Methods, 12:87-98 (1990). Protein renaturation out of GnHCl can be accomplished by several protocols (Jaenicke, R. and Rudolph, R., Protein Structure - A Practical Approach, IRL Press, New York (1990)). Initially, step dialysis is utilized to remove the GnHCL. Alternatively, the purified protein isolated from the Ni-chelate column can be bound to a second column over which a decreasing linear GnHCL gradient is run. The protein is allowed to renature while bound to the column and is subsequently eluted with a buffer containing 250 mM Imidazole, 150 mM NaCl, 25 mM Tris-HCl pH 7.5 and 10% Glycerol. Finally, soluble protein is dialyzed against a storage buffer containing 5 mM Ammonium Bicarbonate. The purified protein was analyzed by SDS-PAGE.

Example 7: Method for determination of mutation of hMLH2 and hMLH3 in hereditary cancer

Isolation of Genomic Clones

[0135] A human genomic P1 library (Genomic Systems, Inc.) was screened by PCR using primers selected for the cDNA sequence of hMLH2 and hMLH3. Two clones were isolated for hMLH2 using primers 5' AAGCTGCTCTGTTAAAAGCG 3' (SEQ ID NO:52) and 5' GCACCAGCATCCAAGGAG 3' (SEQ ID NO:53) and resulting in a 133 bp product. Three clones were isolated for hMLH3, using primers 5' CAACCATGAGACACATCGC 3' (SEQ ID NO:54) and 5' AGGTTAGTGAAGACTCTGTC 3' (SEQ ID NO:55) resulting in a 121 bp product. Genomic clones were nick-translated with digoxigenindeoxy-uridine 5'-triphosphate (Boehringer Manheim), and FISH was performed as described (Johnson, Cg. et al., Methods Cell Biol., 35:73-99 (1991)). Hybridization with the hMLH3 probe were carried out using a vast excess of human cot-1 DNA for specific hybridization to the expressed hMLH3 locus. Chromosomes were counterstained with 4,6-diamino-2-phenylidole

andpropidium iodide, producing a combination of C- and R-bands. Aligned images for precise mapping were obtained using a triple-band filter set (Chroma Technology, Brattleboro, VT) in combination with a cooled charge-coupled device camera (Photometrics, Tucson, AZ) and variable excitation wavelength filters (Johnson, Cv. et al., Genet. Anal: Tech. Appl., 8:75 (1991)). Image collection, analysis and chromosomal fractional length measurements were done suing the ISee Graphical Program System (Inovision Corporation, Durham, NC).

Transcription coupled Translation Mutation Analysis

[0136] For purposes of IVSP analysis the hMLH2 gene was divided into three overlapping segments. The first segment included codons 1 to 500, while the middle segment included codons 270 to 755, and the last segment included codons 485 to the translational termination site at codon 933. The primers for the first segment were 5' GGATCCTAATACGACTCACTATAGGGAGACCACCATGGAACAATTGCCTGC GG 3' (SEQ ID NO:56) and 5' CCTGCTCCACTCATCTGC 3' (SEQ ID NO:57), for the middle segment were 5' GGATCCTAATACGACTCACTATAGGGAGACCACCATGGAAGATATCTTAAA GTTAATCCG 3' (SEQ ID NO:58) and 5' GGCTTCTTCTACTCTATATGG 3' (SEQ ID NO:59), and for the final segment were 5' GGATCCTAATACGACTCACTATAGGGAGACCACCATGGCAGGTCTTGAAAA CTCTTCG 3' (SEQ ID NO:60) and 5' AAAACAAGTCAGTGAATCCTC 3' (SEQ ID NO:61). The primers used for mapping the stop mutation in patient CW all used the same 5' primer as the first segment. The 3' nested primers were:
5' AAGCACATCTGTTTCTGCTG 3' (SEQ ID NO:62) codons 1 to 369;
5' ACGAGTAGATTCCTTTAGGC 3' (SEQ ID NO:63) codons 1 to 290; and
5' CAGAACTGACATGAGAGCC 3' (SEQ ID NO:64) codons 1 to 214.

[0137] For analysis of hMLH3, the hMLH3 cDNA was amplified as a full-length product or as two overlapping segments. The primers for full-length hMLH3 were 5' GGATCCTAATACGACTCACTATAGGGAGACCACCATGGAGCGAGCTGAGAG C 3' (SEQ ID NO:65) and 5' AGGTTAGTGAAGACTCTGTC 3' (SEQ ID NO:66)

(codons 1 to 863). For segment 1, the sense primer was the same as above and the antisense primer was 5' CTGAGGTCTCAGCAGGC 3' (SEQ ID NO:67) (codons 1 to 472). Segment 2 primers were 5' GGATCCTAATACGACTCACTATAGGGAGACCACCATGGTGTCCATTTCCAG ACTGCG 3' (SEQ ID NO:68) and 5' AGGTTAGTGAAGACTCTGTC 3' (SEQ ID NO:69) (codons 415 to 863). Amplifications were done as described below.

[0138] The PCR products contained recognition signals for transcription by T7 RNA polymerase and for the initiation of translation at thei 5' ends. PCR products were used as templates in coupled transcription-translation reactions containing 40 uCi of ^{36}S-methionine (NEN, Dupont). Samples were diluted in SDS sample buffer, and analyzed by electrophoresis on SDS-polyacrylamide gels containing a gradient of 10 to 20% acrylamide. The gels were fixed, treated with EnHance (Dupont), dried and subjected to autoradiography.

RT-PCR and Direct Sequencing of PCR Products

[0139] cDNAs were generated from RNA of lymphoblastoid or tumor cells with Superscript II (Life Technologies). The cDNAs were then used as templates for PCR. The conditions for all amplifications were 35 cycles at 95 C for 30s, 52 C to 62 C for 60 to 120s, and 70 C for 60 to 120s, in buffer. The PCR products were directly sequenced and cloned into the T-tailed cloning vector PCR2000 (Invitrogen) and sequenced with T7 polymerase (United States Biochemical). For the direct sequencing of PCR products, PCR reactions were first phenolchloroform extracted and ethanol precipitated. Templates were directly sequenced using Sequitherm polymerase (Epicentre Technologies) and gamma-^{32}P labelled primers as described by the manufacturer.

Intron/Exon Boundaries and Genomic Analysis of Mutations

[0140] Intron/exon borders were determined by cycle-sequencing P1 clones using gamma-^{32}P end labelled primers and SequiTherm polymerase as described by the manufacturer. The primers used to amplify the hMLH2 exon containing codons 195 to 233 were 5' TTATTTGGCAGAAAAGCAGAG (SEQ ID NO:70) 3' and

5' TTAAAAGACTAACCTCTTGCC 3' (SEQ ID NO:71), which produced a 215 bp product. The product was cycle sequenced using the primer 5' CTGCTGTTATGAACAATATGG 3' (SEQ ID NO:72). The primers used to analyze the genomic deletion of hMLH3 in patient GC were: for the 5' region amplification 5' CAGAAGCAGTTGCAAAGCC 3' (SEQ ID NO:73) and 5' AAACCGTACTCTTCACACAC 3' (SEQ ID NO:74) which produces a 74 bp product containing codons 233 to 257, primers 5' GAGGAAAAGCTTTTGTTGGC 3' (SEQ ID NO:75) and 5' CAGTGGCTGCTGACTGAC 3' (SEQ ID NO:76) which produce a 93 bp product containing the codons 347 to 377, and primers 5' TCCAGAACCAAGAAGGAGC 3' (SEQ ID NO:77) and 5' TGAGGTCTCAGCAGGC 3' (SEQ ID NO:78) which produce a 99 bp product containing the codons 439 to 472 of hMLH3.

TABLE 5

Summary of Mutations in HMLH2 and HMLH3 from patients affected with HNPCC

Sample	Codon	Nucleotides	cDNA Change	Genomic Change	Predicted Coding Change
HMLH2					
CW	233		Skipped Exon	CAG to TAG	GLN to Stop Codon
HMLH3					
MM, NS, TF	20		CGG to CAG	CGG to CAG	ARG to GLN
GC	268 to 669		1,203 bp Deletion	Deletion	In-frame deletion

GCx	268 to 669	1,203 bp Deletion	Deletion	Frameshift, Truncation

Example 8: Expression via Gene Therapy

[0141] Fibroblasts are obtained from a subject by skin biopsy. The resulting tissue is placed in tissue-culture medium and separated into small pieces. Small chunks of the tissue are placed on a wet surface of a tissue culture flask, approximately ten pieces are placed in each flask. The flask is turned upside down, closed tight and left at room temperature over night. After 24 hours at room temperature, the flask is inverted and the chunks of tissue remain fixed to the bottom of the flask and fresh media (e.g., Ham's F12 media, with 10% FBS, penicillin and streptomycin, is added. This is then incubated at 37 C for approximately one week. At this time, fresh media is added and subsequently changed every several days. After an additional two weeks in culture, a monolayer of fibroblasts emerge. The monolayer is trypsinized and scaled into larger flasks.

[0142] pMV-7 (Kirschmeier, P.T. et al, DNA, 7:219-25 (1988) flanked by the long terminal repeats of the Moloney murine sarcoma virus, is digested with EcoRI and HindIII and subsequently treated with calf intestinal phosphatase. The linear vector is fractionated on agarose gel and purified, using glass beads.

[0143] The cDNA encoding a polypeptide of the present invention is amplified using PCR primers which correspond to the 5' and 3' end sequences respectively. The 5' primer containing an EcoRI site and the 3' primer further includes a HindIII site. Equal quantities of the Moloney murine sarcoma virus linear backbone and the amplified EcoRI and HindIII fragment are added together, in the presence of T4 DNA ligase. The resulting mixture is maintained under conditions appropriate for ligation of the two fragments. The ligation mixture is used to transform bacteria HB101, which are then plated onto agar-containing kanamycin for the purpose of confirming that the vector had the gene of interest properly inserted.

[0144] The amphotropic pA317 or GP+am12 packaging cells are grown in tissue culture to confluent density in Dulbecco's Modified Eagles Medium (DMEM) with 10% calf serum (CS), penicillin and streptomycin. The MSV vector containing the gene is then added to the media and the packaging cells are transduced with the vector. The packaging cells now produce infectious viral particles containing the gene (the packaging cells are now referred to as producer cells).

[0145] Fresh media is added to the transduced producer cells, and subsequently, the media is harvested from a 10 cm plate of confluent producer cells. The spent media, containing the infectious viral particles, is filtered through a millipore filter to remove detached producer cells and this media is then used to infect fibroblast cells. Media is removed from a sub-confluent plate of fibroblasts and quickly replaced with the media from the producer cells. This media is removed and replaced with fresh media. If the titer of virus is high, then virtually all fibroblasts will be infected and no selection is required. If the titer is very low, then it is necessary to use a retroviral vector that has a selectable marker, such as neo or his.

[0146] The engineered fibroblasts are then injected into the host, either alone or after having been grown to confluence on cytodex 3 microcarrier beads. The fibroblasts now produce the protein product.

[0147] Numerous modifications and variations of the present invention are possible in light of the above teachings and, therefore, within the scope of the appended claims, the invention may be practiced otherwise than as particularly described.

WHAT IS CLAIMED IS:

1. An isolated polynucleotide comprising a member selected from the group consisting of:

(a) a polynucleotide encoding a polypeptide having the deduced amino acid sequence of SEQ ID NO:2 or a fragment of said polypeptide;

(b) a polynucleotide encoding a polypeptide having the amino acid sequence encoded by the cDNA contained in ATCC Deposit No. 75649;

(c) a polynucleotide encoding a polypeptide having the deduced amino acid sequence of SEQ ID NO:4 or a fragment of said polypeptide;

(d) a polynucleotide encoding a polypeptide having the amino acid sequence encoded by the cDNA contained in ATCC Deposit No. 75651;

(e) a polynucleotide encoding a polypeptide having the deduced amino acid sequence of SEQ ID NO:6 or a fragment of said polypeptide; and

(f) a polynucleotide encoding a polypeptide having the amino acid sequence encoded by the cDNA contained in ATCC Deposit No. 75650.

2. The polynucleotide of Claim 1 wherein the polynucleotide is DNA.

3. The polynucleotide of Claim 1 wherein said polynucleotide encodes a polypeptide having the deduced amino acid sequence of SEQ ID NO:2.

4. The polynucleotide of Claim 1 wherein said polynucleotide encodes a polypeptide having the deduced amino acid sequence of SEQ ID NO:4.

5. The polynucleotide of Claim 1 wherein said polynucleotide encodes a polypeptide having the deduced amino acid sequence of SEQ ID NO:6.

6. The polynucleotide of Claim 1 wherein said polynucleotide encodes a polypeptide encoded by the cDNA of ATCC Deposit No. 75649.

7. The polynucleotide of Claim 1 wherein said polynucleotide encodes a polypeptide encoded by the cDNA of ATCC Deposit No. 75651.

8. The polynucleotide of Claim 1 wherein said polynucleotide encodes a polypeptide encoded by the cDNA of ATCC Deposit No. 75650.

9. A vector containing the polynucleotide of Claim 1.

10. A host cell genetically engineered with the vector of Claim 9.

11. A process for producing a polypeptide comprising expressing from the host cell of Claim 10 the polypeptide encoded by said DNA.

12. A process for producing cells capable of expressing a polypeptide comprising genetically engineering cells with the vector of Claim 9.

13. A polypeptide comprising a member selected from the group consisting of:

(a) a polypeptide having the deduced amino acid sequence of SEQ ID NO:2 and fragments thereof;

(b) a polypeptide encoded by the cDNA of ATCC Deposit No. 75649 and fragments of said polypeptide;

(c) a polypeptide having the deduced amino acid sequence of SEQ ID NO:4 and fragments thereof;

(d) a polypeptide encoded by the cDNA of ATCC Deposit No. 75651 and fragments of said polypeptide;

(e) a polypeptide having the deduced amino acid sequence of SEQ ID NO:6 and fragments thereof; and

(f) a polypeptide encoded by the cDNA of ATCC Deposit No. 75650 and fragments of said polypeptide.

14. An antibody that specifically binds the polypeptide of claim 13.

ABSTRACT OF THE DISCLOSURE

The present invention discloses three human DNA repair proteins and DNA (RNA) encoding such proteins and a proceudre for producing such proteins by recombinant techniques. One of the human DNA repair proteins, hMLH1, has been mapped to chromosome 3 while hMLH2 has been mapped to chromosome 2 and hMLH3 has been mapped to chromosome 7. The polynucleotide sequences of the DNA repair proteins may be used for therapeutic and diagnostic treatments of a hereditary susceptibility to cancer.